10-701/15-781, Fall 2006, Midterm

- There are 7 questions in this exam (11 pages including this cover sheet).
- Questions are not equally difficult.
- If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
- This exam is open book and open notes. Computers, PDAs, cell phones are not allowed.
- You have 1 hour and 20 minutes. Good luck!

Name:
Andrew ID:

Q	Topic	Max. Score	Score
1	Conditional Independece, MLE/MAP, Probability	12	
2	Decision Tree	12	
3	Neural Network and Regression	18	
4	Bias-Variance Decomposition	12	
5	Support Vector Machine	12	
6	Generative vs. Discriminative Classifier	20	
7	Learning Theory	14	
	Total	100	

1 Conditional Independence, MLE/MAP, Probability (12 pts)

1. (4 pts) Show that $\Pr(X, Y|Z) = \Pr(X|Z)\Pr(Y|Z)$ if $\Pr(X|Y, Z) = \Pr(X|Z)$.

$$\Pr(X,Y|Z) = \Pr(X|Y,Z)\Pr(Y|Z) \quad \text{(chain rule)}$$
$$= \Pr(X|Z)\Pr(Y|Z)$$

common mistake: $\Pr(X|Y,Z) = \Pr(X|Z) \Rightarrow X \perp Y$ given Z

$\Rightarrow \Pr(X,Y|Z) = \Pr(X|Z)\Pr(Y|Z)$

the first $\Rightarrow$ does not hold if the equation is not for all possible values of X, Y, Z.

2. (4 pts) If a data point y follows the Poisson distribution with rate parameter θ, then the probability of a single observation y is

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad \text{for } y = 0, 1, 2, \cdots.$$

You are given data points $y_1, \cdots, y_n$ independently drawn from a Poisson distribution with parameter θ. Write down the log-likelihood of the data as a function of θ.

$$\sum_{i=1}^{n}\left(y_i \log\theta - \theta - \log y_i!\right)$$
$$= \left(\sum_{i=1}^{n} y_i\right)\log\theta - n\theta - \log\left(\prod_{i=1}^{n} y_i!\right)$$

3. (4 pts) Suppose that in answering a question in a multiple choice test, an examinee either knows the answer, with probability p, or he guesses with probability $1 - p$. Assume that the probability of answering a question correctly is 1 for an examinee who knows the answer and $1/m$ for the examinee who guesses, where m is the number of multiple choice alternatives. What is the probability that an examinee knew the answer to a question, given that he has correctly answered it?

$$P(\text{Know answer} \mid \text{correct}) = \frac{P(\text{know answer, correct})}{P(\text{correct})}$$
$$= \frac{p}{p + (1-p)\frac{1}{m}}$$

4 Bias-Variance Decomposition (12 pts)

1. (6 pts) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear regression	low/high	low/high
Polynomial regression with degree 3	low/high	low/high
Polynomial regression with degree 10	low/high	low/high

kNN is a regression model think abt affect of models we covered

2. Let $Y = f(X) + \epsilon$, where ϵ has mean zero and variance σ_ϵ^2. In k-nearest neighbor (kNN) regression, the prediction of Y at point x_0 is given by the average of the values Y at the k neighbors closest to x_0.

   (a) (2 pts) Denote the ℓ-nearest neighbor to x_0 by $x_{(\ell)}$ and its corresponding Y value by $y_{(\ell)}$. Write the prediction $\hat{f}(x_0)$ of the kNN regression for x_0 in terms of $y_{(\ell)}, 1 \leq \ell \leq k$.

   $$\hat{f}(x_0) = \frac{1}{k}\sum_{\ell=1}^{k} y_{(\ell)}$$

   (b) (2 pts) What is the behavior of the bias as k increases?

   increase

   (c) (2 pts) What is the behavior of the variance as k increases?

   decrease

5 Support Vector Machine (12 pts)

Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are $(1,1)$ and $(-1,-1)$. The negative examples are $(1,-1)$ and $(-1,1)$.

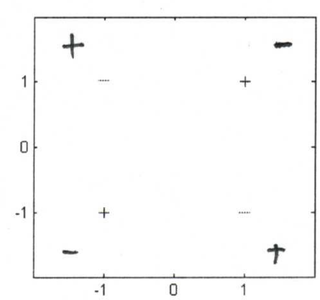
1. (1 pts) Are the positive examples linearly separable from the negative examples in the original space?

   No

2. (4 pts) Consider the feature transformation $\phi(x) = [1, x_1, x_2, x_1 x_2]$, where x_1 and x_2 are, respectively, the first and second coordinates of a generic example x. The prediction function is $y(x) = w^T * \phi(x)$ in this feature space. Give the coefficients, w, of a maximum-margin decision surface separating the positive examples from the negative examples. (You should be able to do this by inspection, without any significant computation.)

   $w = (0,0,0,1)^T$

3. (3 pts) Add one training example to the graph so the total five examples can no longer be linearly separated in the feature space $\phi(x)$ defined in problem 5.2.



4. (4 pts) What kernel $K(x, x')$ does this feature transformation ϕ correspond to?

   $1 + x_1 x_1' + x_2 x_2' + x_1 x_2 x_1' x_2'$

6 Generative vs. Discriminative Classifier (20 pts)

Consider the binary classification problem where class label $Y \in \{0,1\}$ and each training example X has 2 binary attributes $X_1, X_2 \in \{0,1\}$.

In this problem, we will always assume X_1 and X_2 are conditional independent given Y, that the class priors are $P(Y=0) = P(Y=1) = 0.5$, and that the conditional probabilities are as follows:

$P(X_1 \mid Y)$	$X_1 = 0$	$X_1 = 1$
$Y = 0$	0.7	0.3
$Y = 1$	0.2	0.8

$P(X_2 \mid Y)$	$X_2 = 0$	$X_2 = 1$
$Y = 0$	0.9	0.1
$Y = 1$	0.5	0.5

The expected error rate is the probability that a classifier provides an incorrect prediction for an observation: if Y is the true label, let $\hat{Y}(X_1, X_2)$ be the predicted class label, then the expected error rate is

$$P_{\mathcal{D}}\left(Y = 1 - \hat{Y}(X_1, X_2)\right) = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P_{\mathcal{D}}\left(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2)\right).$$

Note that we use the subscript $\mathcal{D}$ to emphasize that the probabilities are computed under the true distribution of the data.

*You don't need to show all the derivation for your answers in this problem.

1. (4 pts) Write down the naïve Bayes prediction for all the 4 possible configurations of X_1, X_2. The following table would help you to complete this problem.

X_1	X_2	$P(X_1, X_2, Y=0)$	$P(X_1, X_2, Y=1)$	$\hat{Y}(X_1, X_2)$
0	0	$0.7 \times 0.9 \times 0.5$	$0.2 \times 0.5 \times 0.5$	0
0	1	$0.7 \times 0.1 \times 0.5$	$0.2 \times 0.5 \times 0.5$	1
1	0	$0.3 \times 0.9 \times 0.5$	$0.8 \times 0.5 \times 0.5$	1
1	1	$0.3 \times 0.1 \times 0.5$	$0.8 \times 0.5 \times 0.5$	1

2. (4 pts) Compute the expected error rate of this naïve Bayes classifier which predicts Y given both of the attributes $\{X_1, X_2\}$. Assume that the classifier is learned with infinite training data.

$0.2 \times 0.5 \times 0.5$
$+ 0.7 \times 0.1 \times 0.5$
$+ 0.3 \times 0.9 \times 0.5$
$+ 0.3 \times 0.1 \times 0.5$
$= 0.235$

3. (4 pts) Which of the following two has a smaller expected error rate?

   - the naïve Bayes classifier which predicts Y given X_1 only
   - the naïve Bayes classifier which predicts Y given X_2 only

Prediction $X_2 = 1 \rightarrow \hat{Y} = 1$ $\quad P_D(X_2=1, Y=0) + P_D(X_2=0, Y=1)$

$X_2 = 0 \rightarrow \hat{Y} = 0$ $\quad = 0.1\times0.5 + 0.5\times0.5 = 0.3$

Prediction

$X_1 = 1 \rightarrow \hat{Y} = 1$

$X_1 = 0 \rightarrow \hat{Y} = 0$

$P_D(X_1=0, Y=0) + P_D(X_1=0, Y=1)$

$= 0.3\times0.5 + 0.2\times0.5 = 0.25$.

4. (4 pts) Now, suppose that we create a new attribute X_3, which is a deterministic copy of X_2. What is the expected error rate of the naïve Bayes which predicts Y given all the attributes (X_1, X_2, X_3) now? Assume that the classifier is learned with infinite training data.

X_1	X_2	$X_3 = X_2$	$P_{NB}(X_1, X_2, X_3, Y=0)$	$P_{NB}(X_1, X_2, X_3, Y=1)$	$\hat{Y}(X_1, X_2, X_3)$
0	0	0	$0.7\times0.9\times0.9\times0.5$	$0.2\times0.5\times0.5\times0.5$	0
0	1	1	$0.7\times0.1\times0.1\times0.5$	$0.2\times0.5\times0.5\times0.5$	1
1	0	0	$0.3\times0.9\times0.9\times0.5$	$0.8\times0.5\times0.5\times0.5$	0
1	1	1	$0.3\times0.1\times0.1\times0.5$	$0.8\times0.5\times0.5\times0.5$	1

Error rate $= 0.2\times0.5\times0.5$ $\quad P_D(X_1=0, X_2=0, Y=1)$

$+ 0.7\times0.1\times0.5$ $\quad P_D(X_1=0, X_2=1, Y=0)$

$+ 0.8\times0.5\times0.5$ $\quad P_D(X_1=1, X_2=0, Y=1)$

$+ 0.3\times0.1\times0.5$ $\quad P_D(X_1=1, X_2=1, Y=0)$

$= 0.3$

5. (4 pts) Explain what is happening with naïve Bayes in problem 6.4? Does logistic regression suffer from the same problem? Why?

The conditional independence assumption of naïve Bayes classifier does not hold. X_2 is overcounted leading to an error prediction when $X_1=0, X_2=0$.

LR does not suffer because it does not make such conditional independence assumption.